Journal of Computer and System Sciences **59**, 392–411 (1999)
Article ID jcss.1999.1635, available online at http://www.idealibrary.com on IDEAL®

Analysis of Two Gradient-Based Algorithms

Nicolò Cesa-Bianchi

DSI, University of Milan, Via Comelico 39, 20135 Milan, Italy
E-mail: cesabian@dsi.unimi.it

Received December 8, 1997; revised March 5, 1999

In this paper we present a new analysis of two algorithms, Gradient Descent and Exponentiated Gradient, for solving regression problems in the on-line framework. Both these algorithms compute a prediction that depends linearly on the current instance, and then update the coefficients of this linear combination according to the gradient of the loss function. However, the two algorithms have distinctive ways of using the gradient information for updating the coefficients. For each algorithm, we show general regression bounds for any convex loss function. Furthermore, we show special bounds for the absolute and the square loss functions, thus extending previous results by Kivinen and Warmuth. In the nonlinear regression case, we show general bounds for pairs of transfer and loss functions satisfying a certain condition. We apply this result to the Hellinger loss and the entropic loss in case of logistic regression (similar results, but only for the entropic loss, were also obtained by Helmbold *et al.* using a different analysis.) Finally, we describe the connection between our approach and a general family of gradient-based algorithms proposed by Warmuth *et al.* in recent works.

1. INTRODUCTION

We study regression problems as an iterated game between a *master predictor* and the *environment*. At each round or *trial* of this game, the master predictor receives from the environment an *instance*, i.e. the values of N real *input variables*, and is challenged to guess an unknown quantity (called *outcome*) also generated by the environment. The master computes its prediction for the outcome by combining the current values of the input variables with the information collected in the past trials. Afterwards, the outcome is revealed and the master incurs a loss computed according to a fixed *loss function*, measuring the discrepancy between the master's prediction and the observed outcome. As no assumptions are made on how the environment generates the sequence of trials, the master could accumulate an

[1] An extended abstract of this paper appeared in the "Proceedings of the 10th Annual Conference on Computational Learning Theory," ACM Press, New York, 1997.

0022-0000/99 $30.00

arbitrarily high loss. To make the model plausible, we adopt a "competitive" approach: The master's goal is to bound, on *any* sequence of trials, the difference between its cumulative loss (i.e. the sum of the losses incurred in each trial) and the corresponding cumulative loss of a "reference predictor," whose predictions are kept hidden from the master.

Using this sequential prediction model, we will show (extending results from [3, 10, 12]) that a well-known algorithm for linear regression, Gradient descent, and a recently proposed variant, Exponentiated Gradient, have a reasonably good performance for a wide range of loss functions even when the regression problem is highly nonlinear and the data are generated with no statistical assumption. As a further motivation for the study of this prediction model, we point out the fact that any good sequential prediction algorithm can be efficiently transformed [2, 12, 15] into an algorithm that performs well in the more traditional statistical (or "batch") frameworks, like those studied in [5, 9].

We use the sequential prediction model to analyze two types of on-line regression problems. In the *linear regression* problem the master algorithm predicts, in each trial t, with a linear combination $\hat{\mathbf{w}}_t \cdot \mathbf{x}_t = \sum_{i=1}^{N} \hat{w}_{t,i} w_{t,i}$ of the input variables $\mathbf{x}_t$, where the coefficients (or weights) $\hat{\mathbf{w}}_t$ of this combination must be chosen independently of $\mathbf{x}_t$. In other words, at the beginning of each trial t we allow the master to modify its choice of weights, but we force it to come up with some choice $\hat{\mathbf{w}}_t$ of weights *before* the current values $\mathbf{x}_t$ of the input variables are revealed. In the *nonlinear regression* problem, the only difference is that the master predicts with the quantity $\phi(\hat{\mathbf{w}}_t \cdot \mathbf{x}_t)$, where ϕ is a fixed and monotone increasing "transfer" function. To match the features of the master predictor, a different reference predictor is associated with each type of regression problem. In linear regression, the reference predictor outputs a linear combination $\bar{\mathbf{w}} \cdot \mathbf{x}_t$ of the input variables, where $\bar{\mathbf{w}}$ are arbitrary but fixed coefficients. Similarly, in nonlinear regression the reference predictor outputs $\phi(\bar{\mathbf{w}} \cdot \mathbf{x}_t)$, where ϕ is the same transfer function used by the master. Our regression bounds will be parametrized by the weights $\bar{\mathbf{w}}$ of the reference predictor. However, none of the master predictors for which we prove these bounds needs the vector $\bar{\mathbf{w}}$ as input parameter.

The two master algorithms we investigate update their weights $\hat{\mathbf{w}}_t$ based on the current gradient of the loss function. One is the classical Gradient descent (GD) algorithm (see [3] for an analysis of this algorithm in the on-line regression framework.) The other is the more recent Exponential Gradient (EG) algorithm [12]. Although both use the gradient of the loss function to update their weights, GD and EG treat this information in two substantially different ways: the former adds to each old weight a term proportional to the current gradient of the loss function, the latter multiplies each old weight by a factor that depends exponentially on the current gradient of the loss function. Master algorithms using similar exponential weights were first used, in related contexts, by Littlestone and Warmuth [17], Feder, Merhav, and Gutman [6], and Vovk [20].

GD and EG have been both analyzed in the on-line regression case with respect to the square loss [3, 12] and in the nonlinear regression case with respect to any pairs of loss and transfer functions satisfying a certain "matching" condition [10]. An example of matching functions are the entropic loss and the logistic transfer

function. In this paper we propose a different analysis of these two gradient-based algorithms. For the linear regression case, in Section 3.2 we show bounds for the square loss (incomparable, though closely related, to those shown by Kivinen and Warmuth in [12]), bounds for the absolute loss (Section 3.1), and general bounds for arbitrary convex loss functions (Section 3). For nonlinear regression, in Section 4 we show general bounds relying on a condition between loss and transfer functions different from the one used by Helmbold, Kivinen, and Warmuth in [10]. We apply this condition to prove bounds for the entropic and Hellinger loss. This latter result was apparently not obtainable with the techniques of [10]. Finally, in Section 5 we briefly describe the connection to some recent work by Warmuth and Jagota [22] and Kivinen and Warmuth [13], where a family of gradient-based algorithms, which includes as special cases the algorithms studied here, is proposed and analyzed.

2. THE PREDICTION GAME

The prediction game we study is parametrized by the number N of input variables and by the loss function L. We call a *loss function* any nonnegative, continuous function L of two real variables y and $\hat{y}$ such that $L(y, \hat{y}) = 0$ whenever $y = \hat{y}$. At the beginning of each trial $t = 1, 2, ...$, the master predictor decides on some vector $\hat{\mathbf{w}}_t$ and receives from the environment the values $\mathbf{x}_t = (x_{t,1}, ..., x_{t,N})$ of the input variables. The master then responds with its prediction $\hat{y}_t = \hat{\mathbf{w}}_t \cdot \mathbf{x}_t$. Finally, the environment decides on an outcome y_t, causing the master to suffer loss $L(y_t, \hat{y}_t)$. We assume that input variables, outcomes, and predictions are all real numbers. Let $\bar{y}_t$ be the prediction of the reference predictor at each trial t. The goal of the master in this prediction game is to minimize the difference $\sum_t L(y_t, \hat{y}_t) - \sum_t L(y_t, \bar{y}_t)$ for an arbitrary sequence of trials (i.e. instance and outcome pairs) chosen by the environment. We prove bounds on this difference that have the general form

$$\sum_{t=1}^{T} L(y_t, \hat{\mathbf{w}}_t \cdot \mathbf{x}_t) - \sum_{t=1}^{T} L(y_t, \bar{\mathbf{w}} \cdot \mathbf{x}_t) \leqslant O(\sqrt{T}) \tag{1}$$

and hold whenever the predictor knows in advance the length T of the trial sequence. Note that bound (1) only applies to those vectors $\bar{\mathbf{w}}$ whose norm is smaller than a constant fixed in advance by the predictor. As this constant also appears in the $O(\sqrt{T})$ term, the bound becomes loose if the best $\bar{\mathbf{w}}$ for the actual trial sequence has norm much smaller than the constant chosen by the predictor.

For the cases where the predictor does not know either the length T of the trial sequence or a good bound on the norm of the best vector $\bar{\mathbf{w}}$, we prove a second set of bounds having form

$$\sum_{t=1}^{T} L(y_t, \hat{\mathbf{w}}_t \cdot \mathbf{x}_t) - \sum_{t=1}^{T} L(y_t, \bar{\mathbf{w}} \cdot \mathbf{x}_t) \leqslant a_\eta \sum_{t=1}^{T} L(y_t, \bar{\mathbf{w}} \cdot \mathbf{x}_t) + b_\eta$$

where a_η and b_η are positive constants depending on a parameter η chosen by the predictor. Although these bounds hold uniformly over T and without any knowledge about the best $\bar{\mathbf{w}}$, they are significantly weaker than (1) whenever $\sum_{t=1}^{T} L(y_t, \bar{\mathbf{w}} \cdot \mathbf{x}_t)$ grows faster than $\sqrt{T}$ for all choices of $\bar{\mathbf{w}}$.

Note that we require that the master's prediction be a linear function of the $\mathbf{x}_t$'s. The situation where the master is allowed to compute $\hat{y}_t$ arbitrarily, that is $\hat{y}_t = \hat{f}_t(\mathbf{x}_t)$ for some arbitrary real function $\hat{f}_t$ (but the reference predictor is still forced to be linear), has been investigated by Vovk [21] and, in a more general framework, by Yamanishi [23]. Vovk's regression bounds have the stronger form

$$\sum_{t=1}^{T} L(y_t, \hat{f}_t(\mathbf{x}_t)) - \sum_{t=1}^{T} L(y_t, \bar{\mathbf{w}} \cdot \mathbf{x}_t) \leqslant \|\bar{\mathbf{w}}\|_2^2 + O(\ln T)$$

and hold for any $\bar{\mathbf{w}}$ and for any trial sequence such that $|y_t|$ is bounded by a constant.

We now describe the Gradient Descent (GD) algorithms and the Exponentiated Gradient (EG) algorithm which will be used as master predictors in the prediction game. Fix the number N of input variables. In each trial t, we will use $\hat{\mathbf{w}}_t$ to denote GD's weight vector and $\hat{\mathbf{p}}_t$ to denote EG's weight vector. GD's prediction for trial t is computed as $\hat{\mathbf{w}}_t \cdot \mathbf{x}_t$ in the linear regression case and as $\phi(\hat{\mathbf{w}}_t \cdot \mathbf{x}_t)$ in the nonlinear regression case, where ϕ is the fixed transfer function. Similarly, EG's predictions for trial t are $\hat{\mathbf{p}}_t \cdot \mathbf{x}_t$ and $\phi(\hat{\mathbf{p}}_t \cdot \mathbf{x}_t)$ in the linear and nonlinear regression case, respectively. We will use $\hat{y}_t$ as a shorthand for the prediction of both master algorithms.

GD uses initial weights $\hat{w}_{1,i} = 0$ for $i = 1, ..., N$. After each outcome y_t is revealed, the weight vector $\hat{\mathbf{w}}_t$ is updated according to the rule

$$\hat{\mathbf{w}}_{t+1} = \hat{\mathbf{w}}_t - \eta L'(y_t, \hat{y}_t)\, \mathbf{x}_t, \tag{2}$$

where $\eta > 0$ is the so-called *learning rate* and $L'(y_t, \hat{y}_t)$ is the derivative $\partial L(y_t, x)/\partial x$ evaluated at $x = \hat{y}_t$.

EG makes initial weight assignments $\hat{p}_{1,i} = 1/N$ for $i = 1, ..., N$ and the new weights $\hat{\mathbf{p}}_{t+1}$ are computed according to the rule

$$\hat{p}_{t+1,i} = \frac{\exp(-\eta L'(y_t, \hat{y}_t)\, x_{t,i})\, \hat{p}_{t,i}}{Z_t} \tag{3}$$

for all $i = 1, ..., N$, where $Z_t = \sum_{j=1}^{N} \exp(-\eta L'(y_t, \hat{y}_t)\, x_{t,j})\, \hat{p}_{t,j}$ and $\eta > 0$ is the learning rate. Note that EG's weight vector $\hat{\mathbf{p}}_t$ satisfies, for all t, $\sum_{i=1}^{N} \hat{p}_{t,i} = 1$ and $0 \leqslant \hat{p}_{t,i} \leqslant 1$ for $i = 1, ..., N$ (equivalently, we say that $\hat{\mathbf{p}}_t$ belongs to the probability simplex). Accordingly to analyze EG we will use a reference predictor whose fixed weight vector is also forced to belong to the probability simplex. Furthermore, when analyzing EG we will always implicitly assume that input variables and outcomes can only take positive values. The extension of EG to the case where negative values and arbitrary linear (rather than just convex) combinations are

allowed can be done via a reduction to the convex case, as shown in [12, 16]. Finally note that, whenever $\hat{y} = \hat{\mathbf{w}} \cdot \mathbf{x}$,

$$\frac{\partial L(y, \hat{y})}{\partial \hat{y}} \mathbf{x} = \left(\frac{\partial L(y, \hat{\mathbf{w}} \cdot \mathbf{x})}{\partial \hat{w}_1}, \ldots, \frac{\partial L(y, \hat{\mathbf{w}} \cdot \mathbf{x})}{\partial \hat{w}_N} \right).$$

So both GD and EG really use the gradient of L to update their weights.

We now extend Kivinen and Warmuth's analysis of gradient-based algorithms in [12] from square loss to more general loss functions. Let $\|\mathbf{v}\|_2 = \sum_{i=1}^N |v_i|$ and $\|\mathbf{v}\|_\infty = \max_i |v_i|$ be, respectively, the 2-norm and the infinity norm for an arbitrary vector $\mathbf{v}$. Furthermore, let $D(\bar{\mathbf{p}} \,\|\, \hat{\mathbf{p}}) = \sum_i \bar{p}_i \ln(\bar{p}_i / \hat{p}_i)$ be the Kullback–Leibler distance between any two real vectors $\bar{\mathbf{p}}$ and $\hat{\mathbf{p}}$ belonging to the probability simplex.

3. LINEAR REGRESSION

For any sequence of T trials and for any master predictor A, we write $L^T(A)$ to denote A's cumulative loss $\sum_{t=1}^T L(y_t, \hat{y}_t)$, where $\hat{y}_t$ is A's prediction at trial t. Similarly, $L^T(\bar{\mathbf{w}}) = \sum_{t=1}^T L(y_t, \bar{\mathbf{w}} \cdot \mathbf{x}_t)$ denotes the cumulative loss of the reference predictor using weights $\bar{\mathbf{w}}$ to compute predictions $\bar{\mathbf{w}} \cdot \mathbf{x}_t$. In case the weights of the reference predictor must belong to the probability simplex (e.g. when the master algorithm is EG), we use $\bar{\mathbf{p}}$ to denote these weights and $L^T(\bar{\mathbf{p}})$ to denote $\sum_{t=1}^T L(y_t, \bar{\mathbf{p}} \cdot \mathbf{x}_t)$. We will use $\bar{y}_t$ for both $\bar{\mathbf{w}} \cdot \mathbf{x}_t$ and $\bar{\mathbf{p}} \cdot \mathbf{x}_t$.

For each one of the two master algorithms GD and EG and for each trial t, we now bound the difference $L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t)$. Let the loss function L be such that, for any fixed $y_t > 0$, $L(y_t, \cdot)$ is twice differentiable with second derivative $L''(y_t, \cdot)$ everywhere nonnegative. Throughout the paper, we will use the notation $L'(y, x) = \partial L(y, x)/\partial x$ and $L''(y, x) = \partial^2 L(y, x)/\partial x^2$.

Fix a real y_t. By applying Taylor's theorem to the function $L(y_t, \cdot)$ we get, for all reals $\hat{y}_t$, $\bar{y}_t$ and for some c between $\hat{y}_t$ and $\bar{y}_t$,

$$L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) = (\hat{y}_t - \bar{y}_t) L'(y_t, \hat{y}_t) - \frac{L''(y_t, c)}{2} (\bar{y}_t - \hat{y}_t)^2 \tag{4}$$

$$\leqslant (\hat{y}_t - \bar{y}_t) L'(y_t, \hat{y}_t). \tag{5}$$

This simple expansion motivates the update rule of both master algorithms when using convex loss functions. For GD, one observes the following.

FACT 1. *Let* $\mathbf{x}$, $\bar{\mathbf{w}}$, *and* $\hat{\mathbf{w}}$ *be real vectors. Then, for any real number* z,

$$z(\hat{\mathbf{w}} \cdot \mathbf{x} - \bar{\mathbf{w}} \cdot \mathbf{x}) = \frac{\|\bar{\mathbf{w}} - \hat{\mathbf{w}}\|_2^2}{2} - \frac{\|\bar{\mathbf{w}} - \hat{\mathbf{w}}'\|_2^2}{2} + z^2 \frac{\|\mathbf{x}\|_2^2}{2},$$

where $\hat{\mathbf{w}}' = \hat{\mathbf{w}} - z\mathbf{x}$.

Now, using (5) and Fact 1 with $z = \eta L'(y_t, \hat{y}_t)$ and $\eta > 0$,

$$L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) \leqslant \frac{\|\bar{\mathbf{w}} - \hat{\mathbf{w}}_t\|_2^2 - \|\bar{\mathbf{w}} - \hat{\mathbf{w}}_{t+1}\|_2^2}{2\eta} + \frac{\eta L'(y_t, \hat{y}_t)^2 \|\mathbf{x}_t\|_2^2}{2}, \tag{6}$$

where $\hat{y}_t = \hat{\mathbf{w}}_t \cdot \mathbf{x}_t$ and $\bar{y}_t = \bar{\mathbf{w}} \cdot \mathbf{x}_t$ for $\bar{\mathbf{w}}$ and $\hat{\mathbf{w}}_t$ arbitrary and $\hat{\mathbf{w}}_{t+1}$ computed from $\hat{\mathbf{w}}_t$ according to GD's update rule (2).

For EG, the next lemma establishes an inequality whose form is similar to the one proven in Fact 1. In fact, both inequalities will turn out to be applications of Taylor's theorem, as discussed in Section 5.

LEMMA 2. *Let* $\mathbf{x}$ *be a real vector with nonnegative components. Let* $\bar{\mathbf{p}}$ *and* $\hat{\mathbf{p}}$ *be any two vectors from the probability simplex. Then, for any real number* z,

$$z(\hat{\mathbf{p}} \cdot \mathbf{x} - \bar{\mathbf{p}} \cdot \mathbf{x}) \leqslant D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}) - D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}') + \frac{z^2}{8} \|\mathbf{x}\|_\infty^2,$$

where $\hat{p}_i = e^{-zx_i}\hat{p}_i / \sum_{j=1}^N e^{-zx_j}\hat{p}_j$ *for each* $i = 1, \ldots, N$.

Proof. In Appendix A. ∎

Thus, using (5) and Lemma 2 with $z = \eta L'(y_t, \hat{y}_t)$ and $\eta > 0$,

$$L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) \leqslant \frac{D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}_t) - D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}_{t+1})}{\eta} + \frac{\eta L'(y_t, \hat{y}_t)^2 \|\mathbf{x}_t\|_\infty^2}{8}, \tag{7}$$

where $\hat{y}_t = \hat{\mathbf{p}}_t \cdot \mathbf{x}_t$ and $\bar{y}_t = \bar{\mathbf{p}} \cdot \mathbf{x}_t$ for any $\bar{\mathbf{p}}$ and $\hat{\mathbf{p}}_t$ from the probability simplex and $\hat{\mathbf{p}}_{t+1}$ computed from $\hat{\mathbf{p}}_t$, according to EG's update rule (3).

For any sequence of trials $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$, we write $\Delta_t(\bar{\mathbf{w}})$ to denote $\|\bar{\mathbf{w}} - \hat{\mathbf{w}}_t\|_2^2/2 - \|\bar{\mathbf{w}} - \hat{\mathbf{w}}_{t+1}\|_2^2/2$ and $\Delta_t(\bar{\mathbf{p}})$ to denote $D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}_t) - D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}_{t+1})$. We will use several times the facts

$$\begin{aligned} \sum_t \Delta_t(\bar{\mathbf{w}}) &\leqslant \|\bar{\mathbf{w}} - \hat{\mathbf{w}}_1\|_2^2/2 \leqslant \|\bar{\mathbf{w}}\|_2^2/2, \\ \sum_t \Delta_t(\bar{\mathbf{p}}) &\leqslant D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}_1) \leqslant \ln N \end{aligned} \tag{8}$$

implied by the telescoping structure of the sums, by the positivity of $\|\cdot\|_2^2$ and $D(\cdot \parallel \cdot)$, and by the facts $\hat{w}_{1,i} = 0$ and $\hat{p}_{1,i} = 1/N$ for $1 \leqslant i \leqslant N$.

The next results shows general bounds on GD and EG with respect to an arbitrary convex loss function. Improved bounds for more specific loss functions will be proven later.

THEOREM 3. *Let* $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_T, y_T)$ *be an arbitrary trial sequence. Let the loss function* L *be such that, for any* $y > 0$, $L(y, \cdot)$ *is convex and twice differentiable. Let* $R_2 \geq \max_t \|\mathbf{x}_t\|_2$, $R_\infty \geq \max_t \|\mathbf{x}_t\|_\infty$, *and* $Z \geq \max_t |L'(y_t, \hat{y}_t)|$. *Then, for any vector* $\bar{\mathbf{w}}$,

$$L^T(\mathrm{GD}) - L^T(\bar{\mathbf{w}}) \leq R_2 Z U \sqrt{T}$$

whenever GD *is run with* $\eta = U/(R_2 Z \sqrt{T})$ *such that* $U \geq \|\bar{\mathbf{w}}\|_2$. *Moreover, for any vector* $\bar{\mathbf{p}}$ *from the probability simplex,*

$$L^T(\mathrm{EG}) - L^T(\bar{\mathbf{p}}) \leq R_\infty Z \sqrt{T \ln(N)/2}$$

whenever EG *is run with* $\eta = \sqrt{(8 \ln N)}/(R_\infty Z \sqrt{T})$.

Proof. To get the first bound, apply (6) to every trial $1 \leq t \leq T$, replacing each $L'(y_t, \hat{y}_t)^2$ with its bound Z^2 and replacing each $\|\mathbf{x}_t\|_2^2$ with its bound R_2^2. Then, sum over trials using the first inequality of (8). Finally, replace $\|\bar{\mathbf{w}}\|_2^2$ with U^2 and replace η with $U/(R_2 Z \sqrt{T})$. The proof of the second bound is very similar; apply (7) to each trial, replacing each $L'(y_t, \hat{y}_t)^2$ with its bound Z^2 and replacing each $\|\mathbf{x}_t\|_\infty^2$ with its bound R_∞^2. Then, sum over trials using the second inequality of (8) and replace η with $\sqrt{(8 \ln N)}/(R_\infty Z \sqrt{T})$. ∎

Note that Theorem 3 can be applied to those trial sequences for which bounds on the quantities $\max_t |L'(y_t, \hat{y}_t)|$ and $\max_t \|\mathbf{x}_t\|_2$ (or $\max_t \|\mathbf{x}_t\|_\infty$) are available in advance for the tuning of η. Instead, the choice of U does not affect the set of trial sequences to which the first bound of the theorem can be applied. To appreciate the influence of U, rewrite the bound as

$$L^T(\mathrm{GD}) \leq \inf_{\|\bar{\mathbf{w}}\|_2 \leq U} L^T(\bar{\mathbf{w}}) + R_2 Z U \sqrt{T}. \tag{9}$$

Now note that the first term in the right-hand side of (9) is clearly nonincreasing in U, whereas the second term increases linearly in U. A similar trade-off, with $\|\bar{\mathbf{w}}\|_\infty$ playing the role of $\|\bar{\mathbf{w}}\|_2$, arises also when EG is replaced by its variant for dealing with arbitrary (rather than just convex) linear combinations (see [12] for examples of this trade-off.)

3.1. *Absolute Loss Bounds*

The absolute loss $L(y, x) = |x - y|$ has first derivative $\partial L(y, x)/\partial x$ not continuous in $x = y$. Hence, (4) is not applicable in this case. However, we can prove bounds similar to those proven for convex functions using the function

$$F(y, x) = \begin{cases} -1 & \text{if } x < y, \\ 0 & \text{if } x = y, \\ 1 & \text{if } x > y, \end{cases} \tag{10}$$

in place of the derivative L'.

THEOREM 4. *Let* $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_T, y_T)$ *be an arbitrary trial sequence. Let* L *be the absolute loss* $L(y, x) = |y - x|$ *and let* GD *and* EG *be run with* F *defined in* (10) *playing the role of* L'. *Let* $R_2 \geqslant \max_t \|\mathbf{x}_t\|_2$ *and* $R_\infty \geqslant \max_t \|\mathbf{x}_t\|_\infty$. *Then, for any vector* $\bar{\mathbf{w}}$,

$$L^T(\mathrm{GD}) - L^T(\bar{\mathbf{w}}) \leqslant R_2 U \sqrt{T}$$

whenever GD *is run with* $\eta = U/(R_2\sqrt{T})$ *such that* $U \geqslant \|\bar{\mathbf{w}}\|_2$. *Moreover, for any vector* $\bar{\mathbf{p}}$ *from the probability simplex,*

$$L^T(\mathrm{EG}) - L^T(\bar{\mathbf{p}}) \leqslant R_\infty \sqrt{T \ln(N)/2}$$

whenever EG *is run with* $\eta = \sqrt{(8 \ln N)}/(R_\infty \sqrt{T})$.

Proof. Note that $L(y, \hat{y}) - L(y, \bar{y}) \leqslant (\hat{y} - \bar{y}) F(y, \hat{y})$ holds for all y, $\hat{y}$, and $\bar{y}$, as one can easily check. Moreover, $F(y_t, \hat{y}_t)^2 = 1$ for all t. Then follow the proof of Theorem 3 with $Z = 1$ and F playing the role of L'. ∎

Note that, for the absolute loss function discussed in Theorem 4, the update rules (2) and (3) for algorithms GD and EG reduce to $\hat{\mathbf{w}}_{t+1} = \hat{\mathbf{w}}_t \pm \eta \mathbf{x}_t$ for GD and to

$$\hat{p}_{t+1,i} = \frac{\exp(\pm \eta x_{t,i})\, \hat{p}_{t,i}}{Z_t}$$

for EG, where the sign is decided according to whether $\hat{y}_t < y_t$ or $\hat{y}_t > y_t$. This special form of EG's update rule is similar to the one used by the Winnow II algorithm described in [14].

Theorem 4 was independently proven by Long [18] (who also shows a matching lower bound) and Bylander [1]. Littlestone and Warmuth [17] prove similar (actually stronger) bounds, but a simpler regression model.

As a final remark, note that Theorem 4 trivially implies a bound for the case where the outcomes y_t all satisfy $0 \leqslant y_t \leqslant 1$ and the reference predictor uses an arbitrary but *constant* prediction $\bar{y}$ chosen from $[0, 1]$. To see this, apply the theorem with $N = 2$, where the two input variables are such that $x_{t,1} = 0$ and $x_{t,2} = 1$ for all t. This extends to the absolute loss (and to real-valued outcomes) recent results by Freund [7] (see also [23] for more general related results.)

3.2. *Square Loss Bounds*

The square loss $L(y, x) = (x - y)^2$ enjoys some nice properties as far as our analysis of gradient-based algorithms is concerned. Because of that, we can prove regression bounds better than those proven in Theorem 3 for general convex functions. These bounds are expressed in terms of additional quantities $\tilde{L}^T(\bar{\mathbf{w}})$ and $\tilde{L}^T(\bar{\mathbf{p}})$ and become better as these quantities become larger.

THEOREM 5. *Let* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ *be an arbitrary trial sequence. Let* L *be the square loss* $L(y, x) = (x - y)^2$. *Let* $R_2 \geq \max_t \|\mathbf{x}_t\|_2$ *and* $R_\infty \geq \max_t \|\mathbf{x}_t\|_\infty$. *Then, for any vector* $\bar{\mathbf{w}}$,

$$L^T(\mathrm{GD}) \leq \frac{1}{1-c}\left(L^T(\bar{\mathbf{w}}) - \tilde{L}^T(\bar{\mathbf{w}}) + \frac{(R_2 \|\bar{\mathbf{w}}\|_2)^2}{c}\right)$$

whenever GD *is run with* $\eta = c/(2R_2^2)$, *where* $0 < c < 1$ *and* $\tilde{L}^T(\bar{\mathbf{w}}) = \sum_{t=1}^T L(\bar{\mathbf{w}} \cdot \mathbf{x}_t, \hat{\mathbf{w}}_t \cdot \mathbf{x}_t)$. *Moreover, for any vector* $\bar{\mathbf{p}}$ *from the probability simplex,*

$$L^T(\mathrm{EG}) \leq \frac{1}{1-c/2}\left(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right)$$

whenever EG *is run with* $\eta = c/R_\infty^2$, *where* $0 < c < 2$ *and* $\tilde{L}^T(\bar{\mathbf{p}}) = \sum_{t=1}^T L(\bar{\mathbf{p}} \cdot \mathbf{x}_t, \hat{\mathbf{p}}_t \cdot \mathbf{x}_t)$.

Proof. We have $L'(y, x) = \partial L(y, x)/\partial x = 2(x - y)$ and $L''(y, x) = \partial^2 L(y, x)/\partial x^2 = 2$. We prove only the bound for EG; the proof for GD is very similar. Let $\hat{y}_t = \hat{\mathbf{p}}_t \cdot \mathbf{x}_t$ and $\bar{y}_t = \bar{\mathbf{p}} \cdot \mathbf{x}_t$. Using (4) and Lemma 2 with $z = \eta L'(y_t, \hat{y}_t)$ and $\eta = c/R_\infty^2$,

$$\begin{aligned} L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) &= L'(y_t, \hat{y}_t)(\hat{y}_t - \bar{y}_t) - (\hat{y}_t - \bar{y}_t)^2 \\ &\leq \frac{\Delta_t(\bar{\mathbf{p}})}{\eta} + \frac{\eta L'(y_t, \hat{y}_t)^2 R_\infty^2}{8} - L(\bar{y}_t, \hat{y}_t) \\ &= \frac{R_\infty^2 \Delta_t(\bar{\mathbf{p}})}{c} + \frac{c}{2} L(y_t, \hat{y}_t) - L(\bar{y}_t, \hat{y}_t). \end{aligned}$$

Now sum over trials to get $L^T(\mathrm{EG}) - L^T(\bar{\mathbf{p}}) \leq (R_\infty^2 \ln N)/c + (c/2) L^T(\mathrm{EG}) - \tilde{L}^T(\bar{\mathbf{p}})$. Under the assumption $c < 2$, we can solve for $L^T(\mathrm{EG})$. Using $D(\bar{\mathbf{p}} \| \hat{\mathbf{p}}_1) \leq \ln N$ we obtain

$$L^T(\mathrm{EG}) \leq \frac{1}{1-c/2}\left(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right),$$

concluding the proof. ∎

Tuning of the parameter c in Theorem 5 is possible if knowledge of a bound on $L^T(\bar{\mathbf{w}}) - \tilde{L}^T(\bar{\mathbf{w}})$ for GD, or a bound on $L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}})$ for EG, is available before the game starts. Lack of this knowledge may be compensated by an iterative scheme obtaining increasingly accurate estimates as the number of observed trials grows (see [3] for details.) Similar remarks hold also for some of the subsequent results.

COROLLARY 6. *For any* $U > 0$ *and any* $\bar{\mathbf{w}}$ *such that* $\|\bar{\mathbf{w}}\|_2 \leq U$, *suppose algorithm* GD *is run with* $\eta = U/(2R_2\sqrt{G})$, *where* $R_2 \geq \max_t \|\mathbf{x}_t\|_2$ *and* $G > \max\{L^T(\bar{\mathbf{w}}) - \tilde{L}^T(\bar{\mathbf{w}}), (R_2 U)^2\}$. *Then,*

$$L^T(\mathrm{GD}) - L^T(\bar{\mathbf{w}}) \leq 2R_2 U\sqrt{G} + 2(R_2 U)^2 - \tilde{L}^T(\bar{\mathbf{w}}) + o(1), \tag{11}$$

where $o(1) \to 0$ *as* $L^T(\bar{\mathbf{w}}) - \tilde{L}^T(\bar{\mathbf{w}}) \to \infty$. *Furthermore, for any* $\bar{\mathbf{p}}$ *from the probability simplex suppose algorithm* EG *is run with* $\eta = \sqrt{2(\ln N)/G'}/R_\infty$, *where* $R_\infty \geqslant \max_t \|\mathbf{x}_t\|_\infty$ *and* $G' > \max\{L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}), (R_\infty^2 \ln N)/2\}$. *Then*

$$L^T(\mathrm{EG}) - L^T(\bar{\mathbf{p}}) \leqslant R_\infty \sqrt{2G' \ln N} + R_\infty^2 \ln N - \tilde{L}^T(\bar{\mathbf{p}}) + o(1), \tag{12}$$

where $o(1) \to 0$ *as* $L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) \to \infty$.

Proof. Again, we just prove the corollary for EG. The setting of η and G' allows us to apply Theorem 5 with $c = R_\infty \sqrt{2(\ln N)/G'} < 2$. Thus we have

$$\begin{aligned} L^T(\mathrm{EG}) &\leqslant \frac{1}{1 - c/2}\left(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right) \\ &= \left(\sum_{k=0}^{\infty} (c/2)^k\right)\left(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right) \\ &= L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c} \\ &\quad + \frac{c(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}))}{2} + \frac{R_\infty^2 \ln N}{2} \\ &\quad + \left(\sum_{k=2}^{\infty} (c/2)^k\right)\left(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right). \end{aligned}$$

By plugging $R_\infty \sqrt{2(\ln N)/G'}$ for c we find that

$$\begin{aligned} L^T(\mathrm{EG}) &\leqslant L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + R_\infty \sqrt{2G' \ln N} + \frac{R_\infty^2 \ln N}{2} \\ &\quad + \left(\sum_{k=2}^{\infty} (c/2)^k\right)\left(L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right) \\ &= L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) + R_\infty \sqrt{2G' \ln N} \\ &\quad + R_\infty^2 \ln N + o(1), \end{aligned}$$

where $o(1) \to 0$ as $L^T(\bar{\mathbf{p}}) - \tilde{L}^T(\bar{\mathbf{p}}) \to \infty$. ∎

Cesa-Bianchi *et al.*'s bounds for GD [3] and Kivinen and Warmuth's bounds fore EG [12] are, respectively, of the form

$$L^T(\mathrm{GD}) - L^T(\bar{\mathbf{w}}) \leqslant 2R_2 U \sqrt{K} + (R_2 U)^2 \tag{13}$$

$$L^T(\mathrm{EG}) - L^T(\bar{\mathbf{p}}) \leqslant R_\infty \sqrt{2K \ln N} + \frac{R_\infty^2 \ln N}{2} \tag{14}$$

for any $K \geqslant L^T(\bar{\mathbf{p}})$. Note that these bounds and the correspond bounds (11) and (12) of Corollary 6 are not comparable. Also, Theorem 7.1 in [3] shows that, for any choice of R_2, U and $K = L^T(\bar{\mathbf{w}})$, and for any master algorithm A, there is a trial sequence such that $L^T(A) - L^T(\bar{\mathbf{w}})$ is at least as big as the right-hand side of (13). This shows that bound (13) is the best possible among all bounds that only depend on the quantities $R_2 = \max_t \|\mathbf{x}_t\|_2$, $U = \|\bar{\mathbf{w}}\|_2$, and $K = L^T(\bar{\mathbf{w}})$. For a careful discussion of square loss lower bounds in terms of quantities R_∞ and $K = L^T(\bar{\mathbf{p}})$, which allow us to measure the tightness of EG upper bounds (12) and (14), the reader is referred to [12].

Results similar to Theorem 5 and Corollary 6 can be proven for loss functions of the form $L(y, x) = |x - y|^p$, where $1 < p < 2$. We have

$$L'(y, x) = \frac{\partial L(y, x)}{\partial x} = \begin{cases} -p(y-x)^{p-1}, & \text{if } x < y, \\ p(x-y)^{p-1}, & \text{if } x \geqslant y. \end{cases}$$

Thus, similarly to the square loss, $L'(y, x)^2 = p^2(x-y)^{2p-2} \leqslant p^2 L(y, x)$ for any reals x, y and for all $1 < p < 2$. Furthermore, $L''(y, x) = \partial^2 L(y, x)/\partial x^2$ is non-negative for all x and y.

THEOREM 7. *Fix any $1 < p < 2$. Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ be an arbitrary trial sequence. Let L be the loss function $L(y, x) = |x - y|^p$. Let $R_2 \geqslant \max_t \|\mathbf{x}_t\|_2$ and $R_\infty \geqslant \max_t \|\mathbf{x}_t\|_\infty$. Then, for any vector $\bar{\mathbf{w}}$,*

$$L^T(\mathrm{Gd}) \leqslant \frac{1}{1 - cp^2/2} \left(L^T(\bar{\mathbf{w}}) + \frac{(R_2 \|\bar{\mathbf{w}}\|_2)^2}{2c} \right)$$

whenever GD *is run with $\eta = c/R_2^2$, where $0 < c < 2/p^2$. Moreover, for any vector $\bar{\mathbf{p}}$ from the probability simplex,*

$$L^T(\mathrm{EG}) \leqslant \frac{1}{1 - cp^2/8} \left(L^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c} \right)$$

whenever EG *is run with $\eta = c/R_\infty^2$, where $0 < c < 8/p^2$.*

Proof. Again, we just give the proof of EG. Using (5) and Lemma 2 with $z = \eta L'(y_t, \hat{y}_t)$ and $\eta = c/R_\infty^2$, we establish the chain of inequalities

$$\begin{aligned} L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) &\leqslant L'(y_t, \hat{y}_t)(\hat{y}_t - \bar{y}_t) \\ &\leqslant \frac{\Delta_t(\bar{\mathbf{p}})}{\eta} + \frac{\eta L'(y_t, \hat{y}_t)^2 R_\infty^2}{8} \\ &\leqslant \frac{R_\infty^2 \Delta_t(\bar{\mathbf{p}})}{c} + \frac{cp^2}{8} L(y_t, \hat{y}_t). \end{aligned}$$

Now sum over trials to get $L^T(\mathrm{EG}) - L^T(\bar{\mathbf{p}}) \leqslant (R_\infty^2 \ln N)/c + (cp^2/8)\, L^T(\mathrm{EG})$. As $c < 8/p^2$, we can solve for $L^T(\mathrm{EG})$ and obtain

$$L^T(\mathrm{EG}) \leqslant \frac{1}{1 - cp^2/8}\left(L^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c} \right),$$

concluding the proof. ∎

Proper tuning of c in Theorem 7 yields bounds similar to those proven in Corollary 6.

4. NONLINEAR REGRESSION

If the loss function L is convex, but such that $L(y, x)$ grows fast as x moves away from y, then Theorem 3 is not very useful, as the factor $\max_t |L'(y_t, \hat{y}_t)|$ can get very big. We can control this growth by applying a squashing "transfer" function to each prediction of the master.

Let ϕ be a real-valued, monotone increasing and differentiable transfer function. Recall that, in the nonlinear case, GD predicts with $\phi(\hat{\mathbf{w}}_t \cdot \mathbf{x}_t)$ and EG predicts with $\phi(\hat{\mathbf{p}}_t \cdot \mathbf{x}_t)$, where $\hat{\mathbf{w}}_t$ and $\hat{\mathbf{p}}_t$ are the corresponding weight vectors. Let $L'(y_t, \phi(\hat{\mu}_t))$ be the derivative $\partial L(y_t, \phi(x))/\partial x$ evaluated at $x = \hat{\mu}_t$, where we use $\hat{\mu}_t$ to denote both $\hat{\mathbf{w}}_t \cdot \mathbf{x}_t$ and $\hat{\mathbf{p}}_t \cdot \mathbf{x}_t$.

The analysis of nonlinear regression is ruled by the interaction between the loss and the transfer functions. Let $\phi(\bar{\mathbf{w}} \cdot \mathbf{x}_t)$ be the prediction at trial t of the reference predictor using weights $\bar{\mathbf{w}}$. We will use $\bar{\mu}_t$ to denote $\bar{\mathbf{w}} \cdot \mathbf{x}_t$. For a fixed y, let $L''(y, \phi(x)) = \partial^2 L(y, \phi(x))/\partial x^2$. By Taylor's theorem, for some c between $\hat{\mu}_t$ and $\bar{\mu}_t$,

$$\begin{aligned} L(y_t, \phi(\hat{\mu}_t)) &= L(y_t, \phi(\bar{\mu})) \\ &= (\hat{\mu} - \bar{\mu}_t)\, L'(y_t, \phi(\hat{\mu}_t)) \\ &\quad - \frac{L''(y_t, \phi(c))}{2} (\hat{\mu}_t - \bar{\mu}_t)^2. \end{aligned} \tag{15}$$

Suppose now L and ϕ "match," so that for some $a > 0$ and for all x and y,

$$L''(y, \phi(x)) \geqslant 0 \tag{16}$$

$$L'(y, \phi(x))^2 \leqslant a L(y, \phi(x)). \tag{17}$$

Then we can derive inequalities similar to (6) and (7). Namely, using (15) and Fact 1 with $z = \eta L'(y_t, \phi(\hat{\mu}_t))$, we have

$$L(y_t, \phi(\hat{\mu}_t)) - L(y_t, \phi(\bar{\mu}_t)) \leqslant \frac{\Delta_t(\bar{\mathbf{w}})}{\eta} + \frac{\eta a L(y_t, \phi(\bar{\mu}_t))\, \|\mathbf{x}_t\|_2^2}{2}. \tag{18}$$

Similarly, using (15) and Lemma 2 with $z = \eta L'(y_t, \phi(\hat{\mu}_t))$, we get

$$L(y_t, \phi(\hat{\mu}_t)) - L(y_t, \phi(\bar{\mu}_t)) \leqslant \frac{\Delta_t(\bar{\mathbf{p}})}{\eta} + \frac{\eta a L(y_t, \phi(\hat{\mu}_t))\, \|\mathbf{x}_t\|_\infty^2}{8}. \tag{19}$$

These inequalities allow us to prove bounds similar to those proven in Theorem 5. When the transfer function ϕ is used, we write $L_\phi^T(A)$ and $L_\phi^T(\bar{\mathbf{w}})$ to denote the cumulative loss after T trials of, respectively, master algorithm A and reference predictor using weights $\bar{\mathbf{w}}$.

THEOREM 8. *Let* $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_T, y_T)$ *be an arbitrary trial sequence. Let* L *be a twice differentiable loss function and* ϕ *a monotone increasing, twice differentiable transfer function such that condition* (16) *is satisfied and condition* (17) *is satisfied for some* $a > 0$. *Let* $R_2 \geqslant \max_t \|\mathbf{x}_t\|_2$ *and* $R_\infty \geqslant \max_t \|\mathbf{x}_t\|_\infty$. *Then, for any vector* $\bar{\mathbf{w}}$,

$$L_\phi^T(\mathrm{GD}) \leqslant \frac{1}{1 - ac/2}\left(L_\phi^T(\bar{\mathbf{w}}) + \frac{(R_2 \|\bar{\mathbf{w}}\|_2)^2}{2c}\right)$$

whenever GD *is run with* $\eta = c/R_2^2$, *where* $0 < c < 2/a$. *Moreover, for any vector* $\bar{\mathbf{p}}$ *from the probability simplex,*

$$L_\phi^T(\mathrm{EG}) \leqslant \frac{1}{1 - ac/8}\left(L_\phi^T(\bar{\mathbf{p}}) + \frac{R_\infty^2 \ln N}{c}\right)$$

whenever EG *is run with* $\eta = c/R_\infty^2$, *where* $0 < c < 8/a$.

Proof. Omitted (similar to the proof of Theorem 5).

Tuning c in Theorem 8 yields improved bounds, similar to those proven in Corollary 6.

COROLLARY 9. *For any vector* $\bar{\mathbf{w}}$ *such that* $\|\bar{\mathbf{w}}\|_2 \leqslant U$, *suppose algorithm* GD *is run with* $\eta = U/(R_2\sqrt{aG})$, *where* $R_2 \geqslant \max_t \|\mathbf{x}_t\|_2$ *and* $G > \max\{L^T(\bar{\mathbf{w}}), a(R_2 U)^2/4\}$. *Then,*

$$L_\phi^T(\mathrm{GD}) - L_\phi^T(\bar{\mathbf{w}}) \leqslant R_2 U\sqrt{aG} + \frac{a(R_2 U)^2}{2} + o(1),$$

where $o(1) \to 0$ *as* $L_\phi^T(\bar{\mathbf{w}}) \to \infty$. *Furthermore, for any* $\bar{\mathbf{p}}$ *from the probability simplex suppose algorithm* EG *is run with* $\eta = \sqrt{(8 \ln N)/(aG')}/R_\infty$, *where* $R_\infty \geqslant \max_t \|\mathbf{x}_t\|_\infty$ *and* $G' > \max\{L^T(\bar{\mathbf{p}}), (aR_\infty^2 \ln N)/8\}$. *Then,*

$$L_\phi^T(\mathrm{EG}) - L_\phi^T(\bar{\mathbf{p}}) \leqslant R_\infty\sqrt{\frac{aG' \ln N}{2}} + \frac{aR_\infty^2 \ln N}{4} + o(1),$$

where $o(1) \to 0$ *as* $L_\phi^T(\bar{\mathbf{p}}) \to \infty$.

Proof. Omitted (similar to the proof of Corollary 6).

Theorem 8 and Corollary 9 can be applied to get bounds with the entropics loss $L(y, x) = y \ln(y/x) + (1 - y) \ln((1 - y)/(1 - x))$. We will use the logistic function $\phi(x) = (1 + e^{-x})^{-1}$ as the transfer function (nonlinear regression with a logistic transfer function is called *logistic regression*). Note that $\phi(x) \in [0, 1]$ and $\phi'(x) = \phi(x)(1 - \phi(x))$ for all reals x.

LEMMA 10. *Let L be the entropic loss and let ϕ be the logistic function. Then for all $x, y \in [0, 1]$*

$$\frac{\partial^2 L(y, \phi(x))}{\partial x^2} \geqslant 0, \quad \left(\frac{\partial L(y, \phi(x))}{\partial x}\right)^2 \leqslant \frac{L(y, \phi(x))}{2}. \tag{20}$$

Proof. We have

$$\frac{\partial L(y, \phi(x))}{\partial x} = \frac{\phi(x) - y}{\phi(x)(1 - \phi(x))} \cdot \phi(x)(1 - \phi(x)) = \phi(x) - y.$$

Hence, the first inequality of (20) holds because ϕ has the first derivative everywhere nonnegative. The second inequality is equivalent to $(x - y)^2 \leqslant L(y, x)/2$ for all $x, y \in [0, 1]$, which is a well-known relation (see, e.g., [4, Lemma 12.6.1, p. 300]). ∎

Therefore, by using Lemma 10 and applying Theorem 8 with $a = 1/2$, we get logistic regression bounds for the entropic loss. Furthermore, application of Corollary 9 yields the bounds

$$L_\phi^T(\mathrm{Gd}) - L_\phi^T(\bar{\mathbf{w}}) \leqslant R_2 U \sqrt{G/2} + \frac{(R_2 U)^2}{4} + o(1) \tag{21}$$

$$L_\phi^T(\mathrm{EG}) - L_\phi^T(\bar{\mathbf{p}}) \leqslant R_\infty \sqrt{\frac{G' \ln N}{4}} + \frac{R_\infty^2 \ln N}{8} + o(1). \tag{22}$$

The techniques developed by Helmbold *et al.* in [10] can be used to obtain bounds similar to (21) and (22). The analysis of nonlinear regression shown here extends those techniques in much the same way the analysis of the linear case in Section 3 extended the techniques of [12].

To show the generality of our approach, we now prove logistic regression bounds for GD and EG applied to the Hellinger loss function

$$H(y, x) = (\sqrt{y} - \sqrt{x})^2 + (\sqrt{1 - y} - \sqrt{1 - x})^2.$$

It appears that the techniques of [10] do not yield any form of regression bounds for this loss function. In fact, they are shown to work only for loss functions L such that, for all y, the expression

$$\frac{\partial L(y, x)}{\partial x} \frac{1}{x - y}$$

is independent of y. It is easy to check that this condition applies to the entropic loss and does not hold for the Hellinger loss.

LEMMA 11. *Let H be the Hellinger loss and let ϕ be the logistic function. Then for all $x, y \in [0, 1]$*

$$\frac{\partial^2 H(y, \phi(x))}{\partial x^2} \geqslant 0, \quad \left(\frac{\partial H(y, \phi(x))}{\partial x}\right)^2 \leqslant \frac{H(y, \phi(x))}{4}. \tag{23}$$

Proof. In Appendix B. ∎

Using Lemma 11 we can apply Theorem 8 and Corollary 9 with $a = 1/4$ to get logistic regression bounds for the Hellinger loss.

5. EXTENSIONS

A new family of gradient-based algorithms for on-line regression, including both GD and EG as special cases, has been recently proposed by Warmuth and Jagota [22] and, for the multidimensional regression case, by Kivinen and Warmuth [13]. (Both of these works have been done independently of ours.) The algorithms of this family, which are called *general additive algorithms* (or GA for short), are parametrized by *weight transformation functions* $\boldsymbol{\psi}: \mathbb{R}^N \to \mathbb{R}^N$ satisfying certain properties that we will explain in a moment. In this section GA will be analyzed (for simplicity just in the linear regression case) using the same tools and terminology we used to analyze GD and EG.

For each fixed weight transformation function $\boldsymbol{\psi}$, algorithm GA($\boldsymbol{\psi}$) maintains a weight vector $\hat{\mathbf{w}}_t$ updated additively following GD's rule (2); that is, $\hat{\mathbf{w}}_{t+1} = \hat{\mathbf{w}}_t - \eta L'(y_t, \hat{y}_t)\,\mathbf{x}_t$. However, unlike GD, the prediction of GA($\boldsymbol{\psi}$) at trial t is $\hat{y}_t = \boldsymbol{\psi}(\hat{\mathbf{w}}_t) \cdot \mathbf{x}_t$. Clearly, when $\boldsymbol{\psi}$ is the identity function, GA($\boldsymbol{\psi}$) reduces to the usual GD algorithm.

The analysis of GA comes naturally from the proof of Theorem 3, once one realizes that Fact 1 and Lemma 2 are both applications of Taylor's theorem. As usual, let L be a twice differentiable loss function with second derivative everywhere nonnegative. We assume that the weight transformation function $\boldsymbol{\psi}$ is monotone increasing and satisfies the key property

$$\nabla F_{\psi}(\bar{\mathbf{w}}, \hat{\mathbf{w}}) = \boldsymbol{\psi}(\hat{\mathbf{w}}) - \boldsymbol{\psi}(\bar{\mathbf{w}}) \tag{24}$$

for some *weight distance function* $F_{\psi}: \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ and for all $\hat{\mathbf{w}}$ and $\bar{\mathbf{w}}$, where

$$\nabla F_{\psi}(\bar{\mathbf{w}}, \hat{\mathbf{w}}) = \left(\frac{\partial F_{\psi}(\bar{\mathbf{w}}, \hat{\mathbf{w}})}{\partial \hat{w}_1}, \ldots, \frac{\partial F_{\psi}(\bar{\mathbf{w}}, \hat{\mathbf{w}})}{\partial \hat{w}_N} \right).$$

We use $\hat{\boldsymbol{\psi}}_t$ and $\bar{\boldsymbol{\psi}}$ to denote, respectively, $\boldsymbol{\psi}(\hat{\mathbf{w}}_t)$ and $\boldsymbol{\psi}(\bar{\mathbf{w}})$. Also, $\bar{y}_t$ will denote $\bar{\boldsymbol{\psi}} \cdot \mathbf{x}_t$. Applying Taylor's theorem in the same way we did for proving (4) and (5), we find that

$$\begin{aligned} L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) &= (\hat{y}_t - \bar{y}_t)\, L'(y_t, \hat{y}_t) - \frac{L''(y_t, c)}{2} (\bar{y}_t - \hat{y}_t)^2 \\ &\leqslant (\hat{y}_t - \bar{y}_t)\, L'(y_t, \hat{y}_t) \qquad (25) \\ &= (\hat{\boldsymbol{\psi}}_t - \bar{\boldsymbol{\psi}}) \cdot \mathbf{x}_t L'(y_t, \hat{y}_t) \\ &= \frac{1}{\eta} (\hat{\boldsymbol{\psi}}_t - \bar{\boldsymbol{\psi}}) \cdot (\hat{\mathbf{w}}_t - \hat{\mathbf{w}}_{t+1}), \qquad (26) \end{aligned}$$

where we used convexity of L in (25) and invoked GA's update rule (2) for proving (26).

Applying Taylor's theorem once more, this time to the function $F_\psi(\bar{\mathbf{w}}, \cdot)$, we get

$$(\hat{\mathbf{w}}_t - \hat{\mathbf{w}}_{t+1}) \cdot \nabla F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) = F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) - F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_{t+1}) + E_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t, \mathbf{x}_t), \tag{27}$$

where $E_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t, \mathbf{x}_t)$ is the second-order error term in the Taylor expansion of $F_\psi(\bar{\mathbf{w}}, \cdot)$ around $\hat{\mathbf{w}}_{t+1}$. Using (24) to connect (26) with (27), we find that

$$L(y_t, \hat{y}_t) - L(y_t, \bar{y}_t) \leqslant \frac{1}{\eta} [F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) - F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_{t+1})] + \frac{1}{\eta} E_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t, \mathbf{x}_t). \tag{28}$$

Summing up over t we then get the general bound

$$L^T(\mathrm{GA}) - L^T(\bar{\mathbf{w}}) \leqslant \frac{1}{\eta} [F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_1) - F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_{T+1})] + \frac{1}{\eta} \sum_{t=1}^{T} E_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t, \mathbf{x}_t). \tag{29}$$

We now recover the basic bounds for GD and EG shown in Section 3. First, note that (24) holds when $F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) = \frac{1}{2} \|\bar{\mathbf{w}}_t\|_2^2$ and $\boldsymbol{\psi}$ is the identity function. In this case $\mathrm{GA}(\boldsymbol{\psi}) = \mathrm{GD}$ and $E_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t, \mathbf{x}_t) = \frac{1}{2}\eta^2 L'(y_t, \hat{y}_t)^2 \|\mathbf{x}_t\|_2^2$ by Fact 1; thus (28) reduces to (6) shown in Section 3. Second, (24) holds also when

$$\psi(\mathbf{w})_i = \frac{e^{w_i}}{\sum_{j=1}^N e^{w_j}} \qquad \text{for} \quad i = 1, \ldots, N;$$

$$F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) = D(\bar{\boldsymbol{\psi}} \,\|\, \hat{\boldsymbol{\psi}}_t).$$

In this second case $\mathrm{GA}(\boldsymbol{\psi}) = \mathrm{EG}$ and $E_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t, \mathbf{x}_t) \leqslant \frac{1}{8}\eta^2 L'(y_t, \hat{y}_t)^2 \|\mathbf{x}_t\|_\infty^2$ by Lemma 2; thus (28) reduces to (7), also shown in Section 3.

As suggested in [13], we can nicely express the error term E_ψ by choosing

$$F_\psi(\mathbf{v}, \mathbf{w}) = \int_{\mathbf{v}}^{\mathbf{w}} (\boldsymbol{\psi}(\mathbf{u}) - \boldsymbol{\psi}(\mathbf{v}))\, d\mathbf{u}$$

whenever $\boldsymbol{\psi}$ is such that the value of the integral does not depend on the path taken from $\mathbf{v}$ to $\mathbf{w}$. This choice of F_ψ clearly satisfies (24). Moreover, one can prove that

$$(\hat{\mathbf{w}}_t - \hat{\mathbf{w}}_{t+1})\, \nabla F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) = F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_t) - F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_{t+1}) + F_\psi(\hat{\mathbf{w}}_t, \hat{\mathbf{w}}_{t+1}) \tag{30}$$

for all $\bar{\mathbf{w}}$, $\hat{\mathbf{w}}_t$, and $\hat{\mathbf{w}}_{t+1}$. Doing the derivation of (29) again, this time using (30) instead of (27), we then get the appealing form

$$L^T(\mathrm{GA}) - L^T(\bar{\mathbf{w}}) \leqslant \frac{1}{\eta}\left[F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_1) - F_\psi(\bar{\mathbf{w}}, \hat{\mathbf{w}}_{T+1})\right] + \frac{1}{\eta}\sum_{t=1}^{T} F_\psi(\hat{\mathbf{w}}_t, \hat{\mathbf{w}}_{t+1})$$

originally shown in Theorem 1 of [22].

We close by observing that the notions of weight transformation and weight distance functions have been also used by Grove *et al.* in [8], where they have proposed a broad family of algorithms, similar to GA, for solving on-line binary classification problems. Notwithstanding the generality of Grove *et al.*'s approach, it is not yet clear how to extend their results to the regression framework.

APPENDIX A: PROOF OF LEMMA 2

We use a simple bound (based on convexity arguments) on the log of the moment-generating function of a random variable. This result was originally proven by W. Hoeffding to show bounds on the tails of the binomial distribution. Here we give a proof due to Pollard [19].

LEMMA 12 [11]. *Let Y be a random variable with 0 mean and range $[a, b]$. Then, for any real number z,*

$$\ln \mathbf{E}[e^{zY}] \leqslant \frac{z^2}{8}(b-a)^2.$$

By convexity of the exponential function we have

$$e^{zY} \leqslant \frac{b-Y}{b-a}e^{za} + \frac{Y-a}{b-a}e^{zb}.$$

Taking expectations, recalling that $\mathbf{E}[Y] = 0$, we get

$$\mathbf{E}[e^{zY}] \leqslant \frac{b}{b-a}e^{za} - \frac{a}{b-a}e^{zb} = e^{za}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{z(b-a)}\right).$$

Setting $u = z(b-a)$ and taking logs on both sides yields

$$\ln \mathbf{E}[e^{zY}] \leqslant \frac{a}{b-a}u + \ln\left(\frac{b}{b-a} - \frac{a}{b-a}e^{u}\right).$$

Let $\alpha = -a/(b-a)$. Note that $b/(b-a) = 1-\alpha$. Set $F(u) = -\alpha u + \ln(1-\alpha+\alpha e^{\mathbf{u}})$. Differentiate function F with respect to u:

$$F'(u) = -\alpha + \frac{\alpha e^u}{1-\alpha+\alpha e^u} = -\alpha + \frac{\alpha}{\alpha + (1-\alpha)e^{-u}}$$

$$F''(u) = \frac{\alpha(1-\alpha)e^{-u}}{(\alpha+(1-\alpha)e^{-u})^2} = \frac{\alpha}{\alpha+(1-\alpha)e^{-u}} \cdot \frac{(1-\alpha)e^{-u}}{\alpha+(1-\alpha)e^{-u}} \leqslant \frac{1}{4},$$

where the last inequality holds because $x(1-x) \leqslant 1/4$ for all $0 < x < 1$ and because $\alpha > 0$ since $a < 0 < b$. Applying Taylor's theorem we finally get

$$F(u) = F(0) + uF'(0) + \frac{u^2}{2} F''(v) = \frac{u^2}{2} F''(v) \leqslant \frac{z^2}{8} (b-a)^2,$$

concluding the proof. ∎

Proof of Lemma 2.

$$\begin{aligned} D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}) - D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}') &= \sum_{i=1}^{N} \bar{p}_i \ln \frac{\hat{p}'_i}{\hat{p}_i} \\ &= \sum_{i=1}^{N} \bar{p}_i \ln e^{-zx_i} - \sum_{i=1}^{N} \bar{p}_i \ln \left(\sum_{j=1}^{N} e^{-zx_j} \hat{p}_j \right) \\ &= -z\bar{\mathbf{p}} \cdot \mathbf{x} - \ln \left(\sum_{i=1}^{N} e^{-zx_i} \hat{p}_i \right) \\ &= -z\bar{\mathbf{p}} \cdot \mathbf{x} - \ln \left[\sum_{i=1}^{N} \exp(-zx_i + z\hat{\mathbf{p}} \cdot \mathbf{x} - z\hat{\mathbf{p}} \cdot \mathbf{x})\, \hat{p}_i \right] \\ &= -z\bar{\mathbf{p}} \cdot \mathbf{x} + z\hat{\mathbf{p}} \cdot \mathbf{x} - \ln \left(\sum_{i=1}^{N} e^{-zv_i} \hat{p}_i \right), \end{aligned}$$

where $v_i = x_i - \hat{\mathbf{p}} \cdot \mathbf{x}$. Hence, $z(\hat{\mathbf{p}} \cdot \mathbf{x} - \bar{\mathbf{p}} \cdot \mathbf{x}) = D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}) - D(\bar{\mathbf{p}} \parallel \hat{\mathbf{p}}') + \ln(\sum_{i=1}^{N} e^{-zv_i} \hat{p}_i)$. Using Lemma 12 and the assumption on $\mathbf{x}$, we find that $\ln(\sum_{i=1}^{N} e^{-zv_i} \hat{p}_i) \leqslant (z \|\mathbf{x}\|_\infty)^2/8$, concluding the proof. ∎

APPENDIX B: PROOF OF LEMMA 11

We have

$$\begin{aligned} \frac{\partial H(y,z)}{\partial s} &= \frac{\sqrt{1-y} - \sqrt{1-z}}{\sqrt{1-z}} - \frac{\sqrt{y} - \sqrt{z}}{\sqrt{z}} = \sqrt{\frac{1-y}{1-z}} - \sqrt{\frac{y}{z}}, \\ \frac{\partial^2 H(y,z)}{\partial z^2} &= \frac{1}{2} \sqrt{\frac{1-y}{(1-z)^3}} + \frac{1}{2} \sqrt{\frac{y}{z^3}}. \end{aligned}$$

Furthermore, $\phi' = \phi \cdot (1-\phi)$ and $\phi'' = \phi' \cdot (1-2\phi)$. As

$$\frac{\partial H(y, \phi(x))}{\partial x} = \frac{\partial H(y,\phi)}{\partial \phi} \cdot \phi'$$

and

$$\frac{\partial^2 H(y, \phi(x))}{\partial x^2} = \frac{\partial^2 H(y,\phi)}{\partial \phi^2} \cdot \phi' + \frac{\partial H(y,\phi)}{\partial \phi} \cdot \phi''$$

to prove the first inequality of (23) we just have to show

$$\frac{\partial^2 H(y,\phi)}{\partial \phi^2} + \frac{\partial H(y,\phi)}{\partial \phi} \cdot (1-2\phi) \geqslant 0.$$

Or, equivalently,

$$\frac{1}{2}\sqrt{\frac{1-y}{(1-\phi)^3}} + \frac{1}{2}\sqrt{\frac{y}{\phi^3}} + \left(\sqrt{\frac{1-y}{1-\phi}} - \sqrt{\frac{y}{\phi}}\right) \cdot (1-2\phi) \geqslant 0.$$

Collecting terms we find that

$$\sqrt{\frac{1-y}{1-\phi}}\left(\frac{1}{2(1-\phi)} + 1 - 2\phi\right) + \sqrt{\frac{y}{\phi}}\left(\frac{1}{2\phi} - 1 + 2\phi\right) \geqslant 0.$$

Finally, reducing to common denominator each term in the left-hand side and multiplying both sides by 2 we get

$$\sqrt{\frac{1-y}{1-\phi}} \cdot \frac{4\phi^2 - 6\phi + 3}{1-\phi} + \sqrt{\frac{y}{\phi}} \cdot \frac{4\phi^2 - 2\phi + 1}{\phi} \geqslant 0.$$

It is easily verified that $4\phi^2 - 6\phi + 3$ and $4\phi^2 - 2\phi + 1$ are both always positive. This concludes the proof of the first inequality of (23).

Now let $A = \sqrt{y} - \sqrt{\phi}$ and $B = \sqrt{1-y} - \sqrt{1-\phi}$. Then the second inequality in (23) is equivalent to

$$\left(\frac{B}{\sqrt{1-\phi}} - \frac{A}{\sqrt{\phi}}\right)^2 \phi^2 (1-\phi) \leqslant \frac{A^2 + B^2}{4}.$$

As $\phi(1-\phi) \leqslant 1/4$, the above is, in turn, implied by

$$\left(\frac{B}{\sqrt{1-\phi}} - \frac{A}{\sqrt{\phi}}\right)^2 \phi(1-\phi) \leqslant A^2 + B^2.$$

Expanding the square and multiplying through in the left-hand side yields

$$B^2\phi + A^2(1-\phi) - 2AB\sqrt{\phi(1-\phi)} \leqslant A^2 + B^2$$

which is easily seen to hold. This concludes the proof of the lemma. ∎

ACKNOWLEDGMENTS

The author acknowledges support of ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II).

REFERENCES

1. T. Bylander, Worst-case absolute loss bounds for linear learning algorithms, *in* "Proceedings of the 14th National conference on Artificial Intelligence," pp. 485–490, MIT Press, Cambridge, MA, 1997.
2. N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Shapire, and M. K. Warmuth, How to use expert advice, *J. Assoc. Compute Mach.* **44**, No. 3 (1997), 427–485.
3. N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent, *IEEE Trans. Neural Networks* **7**, No. 3 (1996), 604–619.
4. T. M. Cover and J. A. Thomas, "Elements of Information Theory," Wiley, New York, 1991.
5. L. Devroye, L. Győrfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," Springer-Verlag, New York/Berlin, 1996.
6. M. Feder, N. Merhav, and M. Gutman, Universal prediction of individual sequences, *IEEE Trans. Inform. Theory* **38** (1992), 1258–1270.
7. Y. Freund, Predicting a binary sequence almost as well as the optimal biased coin, *in* "Proceedings of the 9th Annual Conference on Computational Learning Theory," pp. 89–98, ACM Press, New York, 1996.
8. A. J. Grove, N. Littlestone, and D. Schuurmans, General convergence results for linear discriminant updates, *in* "Proceedings of the 10th Annual Conference on Computational Learning Theory," pp. 171–183, ACM Press, 1997.
9. D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. and Comput.* **100**, No. 1 (1992), 78–150.
10. D. P. Helmbold, J. Kivinen, and M. K. Warmuth, Worst-case loss bounds for sigmoided neurons, *in* "Advances in Neural Information Processing Systems 8," pp. 309–315, MIT Press, 1997.
11. W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58** (1963), 13–30.
12. J. Kivinen and M. K. Warmuth, Exponentiated gradient versus gradient descent for linear predictors, *Inform. and Comput.* **132**, No. 1 (1997), 1–63.
13. J. Kivinen and M. K. Warmuth, Relative loss bounds for multidimensional regression problems, *in* "Advances in Neural Information Processing Systems 9," pp. 287–293, MIT Press, 1998.
14. N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear threshold algorithm, *Mach. Learning* **2**, No. 4 (1988), 285–318.
15. N. Littlestone, From on-line to batch learning, *in* "Proceedings of the 2nd Annual Workshop on Computational Learning Theory," pp. 269–284, Morgan Kaufmann, San Mateo, CA, 1989.
16. N. Littlestone, P. M. Long, and M. K. Warmuth, On-line learning of linear functions, *Comput. Complexity* **5**, No. 1 (1995), 1–23.
17. N. Littlestone and M. K. Warmuth, The weighted majority algorithm, *Inform. and Comput.* **108** (1994), 212–261.
18. P. M. Long, "Absolute Loss Bounds for Prediction Using Linear Functions," Technical Report, TRB7 Dept. of Information Systems and Computer Science, National University of Singapore, 1996.
19. D. Pollard, "Convergence of Stochastic Processes," Springer-Verlag, New York/Berlin, 1984.
20. V. G. Vovk, A game of prediction with expert advice, *Journal of Computer and System Sciences* **56**, No. 2 (1998), 153–173.
21. V. G. Vovk, Competitive on-line linear regression, *in* "Advances in Neural Information Processing Systems 10," pp. 364–370, MIT Press, Cambridge, MA, 1998.
22. M. K. Warmuth and A. K. Jagota, Continuous versus discrete-time nonlinear gradient descent: Relative loss bounds and convergence, *in* "Proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics," 1997.
23. K. Yamanishi, A decision-theoretic extension of stochastic complexity and its application to learning, *IEEE Trans. Inform. Theory* **44** (1998), 1424–1440.